December 2020

Response to European Digital Media Observatory Call for Comments

The GDPR and Sharing Data for Independent Social Scientific Research

FACEBOOK

FACEBOOK

December 21, 2020

To the European Digital Media Observatory:

Facebook welcomes the opportunity to submit comments on your important initiative to launch a Working Group on "Access to Data Held by Digital Platforms for the Purposes of Social Scientific Research."

Facebook supports independent research and unlocking the power of data to solve some of the world's greatest challenges. We are also deeply committed to protecting our users' privacy and maintaining a safe and secure community. We provide these comments in the hopes of better enabling researchers to study the impact of technology on society.

As you rightly note in your proposal, unlocking access to data held by digital platforms for the purposes of independent social scientific research will depend on developing high standards for preserving privacy and clear mechanisms that hold all parties accountable for their access to and use of data. At Facebook, we have worked to promote research—while preserving privacy—through multiple initiatives. For example, our Data For Good program helps researchers and humanitarian organizations respond to emergencies by sharing data, subject to privacy preserving methods like aggregation and de-identification. Likewise, the Facebook Open Research and Transparency Team collaborates with external academics to understand the impact of our platforms on elections and democracy, including the spread of disinformation.

In order to do our part in facilitating social scientific research about issues of public interest, it is vital that we are able to comply with varying privacy frameworks globally, including the European Union's General Data Protection Regulation ("GDPR"), the application of which has proved to be particularly challenging for designing and implementing independent research projects. In the attached comments, we explain some of the legal challenges we have observed, including with regard to: establishing a legal basis, providing adequate transparency, implementing suitable safeguards, addressing overlapping Member State requirements, and de-identifying data to sufficient standards, all while preserving the independence of researchers and the utility of data for research purposes.

A code of conduct, such as the one EDMO is proposing, could help in addressing many of these issues. In particular, a code of conduct would provide clear, credible and enforceable standards for sharing data for research, with the added legal certainty of having such standards approved by data

protection regulators. In so doing, the code would help facilitate more collaborations with academic researchers that are in compliance with GDPR requirements. Additionally, by inviting relevant stakeholders to participate in the development of a code of conduct, EDMO's proposal can ensure that the interests of all affected parties, including data subjects, are adequately protected. In the attached comments, we also outline what we consider to be the key issues to be resolved by a code of conduct for research.

Thank you again for the opportunity to submit our comments on your proposal. We hope these comments provide a useful view into the challenges faced by companies like Facebook and help move this important initiative forward.

Should our participation in the EDMO process or follow-on discussions be of value, we are at your disposal.

Sincerely,

Andrew Gruen, Ph.D.
Facebook Open Research
and Transparency

Hershel S. Eisenberger,
Privacy & Data Policy

TABLE OF CONTENTS

I. INTRODUCTION 1

II. LEGAL BACKGROUND 2

1. Promotion of research in EU law 2
2. GDPR research exemptions 3
3. Using data for independent research 6

III. DATA SHARING CHALLENGES UNDER THE GDPR 9

1. The GDPR's analytical framework discourages data originators from sharing personal data for independent purposes. 9

   a. The GDPR's legal bases for processing do not easily align with sharing data for research purposes. 9

   b. Although researchers may benefit from more relaxed transparency requirements, it is not clear that data originators can also take advantage of these provisions. 16

   c. Data originators are not well-positioned to meet the GDPR's accountability requirements where they cannot analyze the risks of data use by independent parties. 18

2. The GDPR's research exemptions fail to provide sufficient legal assurances for data originators. 20

   a. What type of research is qualifying research under the GDPR is unclear. 20

   b. The purpose limitation exemption for qualifying research requires a careful assessment of appropriate safeguards. 21

3. Variation in Member State law exacerbates risks for cross-border research. 23

a. Key research provisions are not harmonized across the EU. 23

b. Overlapping Member State implementations of the GDPR and other local legal frameworks could apply to cross-border research. 25

4. The standards for effective anonymization under the GDPR remain unsettled. 26

a. Guidance from EU regulators adopts strict and inconsistent standards of anonymization. 26

b. Data originators cannot effectively assess the risks of identification where data will be made available to independent researchers. 29

IV. HOW EDMO'S PROPOSED CODE OF CONDUCT COULD SUPPORT DATA SHARING FOR INDEPENDENT RESEARCH 31

1. Potential benefits of developing a code of conduct. 31

2. Proposed issues to be addressed and relevant stakeholders. 33

V. CONCLUSION 35

The GDPR and Sharing Data for Independent Social Scientific Research

I. INTRODUCTION

This paper analyzes how the European Union's General Data Protection Regulation[1] ("**GDPR**") affects the sharing of personal data held by commercial parties such as Facebook for independent social scientific research. This paper also explains why the GDPR's regime governing research can make it more difficult for companies to share data with independent researchers. As explored in detail in this paper, in the face of ambiguity in the application of core provisions in the context of independent research, commercial entities may refrain from sharing data to minimize legal risks, which in turn may undermine societally beneficial independent social scientific research. Here, we identify the aspects of the GDPR that could hamper efforts to promote sharing data for independent research.[2] This paper concludes by examining the ways in which a code of conduct as proposed by the European Digital Media Observatory ("**EDMO**") could address many of these challenges.

[1] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ 2016 L 119.

[2] For the purpose of this analysis, "independent research" means research by scholars unaffiliated with the commercial party that supplies the data, and where the commercial party does not pre-select the research topics or influence the results of the research.

II. LEGAL BACKGROUND

The promotion of research and the protection of personal data are both core and longstanding objectives of European Union ("**EU**") law. Growing overlap between the two objectives has highlighted points of tension between competing values and requirements under EU law. These tensions, and particularly, increasing emphasis on personal data protection under EU law, has led commercial parties such as Facebook to take a cautious approach to sharing data for research purposes.

1. Promotion of research in EU law

The promotion of research is an important feature of the EU political project. In 1957, the Treaty of Rome, which created the European Economic Community ("**EEC**"), empowered the EEC to develop measures for the "effective co-ordination of efforts in the spheres of vocational training, of research and of the dissemination of agricultural knowledge."[3] This mandate has expanded along with the spheres of activity of the EU to include the promotion of "a European research area in which researchers, scientific knowledge and technology circulate freely."[4] In recent years, the European Commission's "Digital Single Market Strategy for Europe"[5], "European Data Strategy"[6] and "Common European Data Spaces"[7] initiatives have prioritized research—and particularly big data research—within the EU, and emphasized the need for sharing of such data to promote the competitiveness of EU industry and the well-being of EU residents. Recognizing the close connection between data sharing, research and data protection considerations, the European Commission and the High

[3] EEC Treaty, art. 41(a) (as in effect 1958).

[4] TFEU Treaty, art. 179 (as in effect 2009).

[5] European Commission, *A Digital Single Market Strategy for Europe*, COM (2015) 192 final (May 6, 2015). The strategy sought to maximize "[C]loud computing and Big Data, and research and innovation," and called for "greater legal certainty" for researchers and greater "access to public data to help drive innovation." *Id.*

[6] European Commission, *A European strategy for data*, COM (2020) 66 final (Feb. 19, 2020).

[7] European Commission, *Data sharing in the EU - Common European Data Spaces* (Feb 2020-June 2020).

Representative's "Joint Communication on Tackling COVID 19 disinformation" specifically invited EDMO to collaborate on developing "a framework providing academic researchers privacy-protected access to relevant platforms' data to enhance the detection and analysis of disinformation."[8]

2. GDPR research exemptions

At the same time as the EU has sought to promote social scientific and digital research, it has also pushed for a stronger and more modern framework for protecting personal data. In 2016, the EU enacted the GDPR, which went into effect in May 2018.[9] Positioned as a pillar of the European Commission's "Digital Single Market Strategy for Europe,"[10] the GDPR expanded upon data protection principles that had been enshrined in the earlier Data Protection Directive from 1995 (the "**Directive**"), while modernizing many of its provisions to address the rise of digital technologies.

Recognizing the potential for these rules to grate against research objectives, the GDPR offered a more flexible regime for "scientific or historical research purposes or statistical purposes" (collectively, "**qualifying research**"),[11] intended to promote the development of a "European research area" in accordance with EU law.[12] Although research is not exempt from GDPR requirements altogether,[13] the GDPR encourages a balanced approach that would facilitate qualifying research by relaxing several key requirements, as long as "appropriate safeguards" are put in place to protect individuals (the "**research exemptions**").[14]

---

[8] Joint Communication of the European Commission and the High Representative for Foreign Affairs and Security Policy, *Tackling COVID-19 disinformation - Getting the facts right*. JOIN(2020) 8 final (June 10, 2020).

[9] Giovanni Buttarelli, *The EU GDPR as a clarion call for a new global digital gold standard*, European Data Protection Supervisor (Apr. 1, 2016).

[10] *European Commission Proposal for a Digital Single Market Strategy for Europe*, COM(2015) 192 final (May 6, 2020).

[11] GDPR, Article 89.

[12] GDPR, Recital 159.

[13] *Id.* ("Where personal data are processed for scientific research purposes, this Regulation should also apply to that processing.").

[14] GDPR, Art. 89(1).

In broad strokes, the research exemptions loosen three categories of requirements, which, without such exemptions, could impair research objectives. First, the research compatibility exemption permits controllers to presume that using personal data for research purposes is "compatible" with the initial purposes of data collection.[15] This is critical for addressing the GDPR's purpose limitation principle, which stipulates that personal data must be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes."[16] Without this exemption, companies could be concerned about using data collected for one purpose—for example, in connection with providing services—for subsequent research purposes without establishing an additional legal basis (or without a detailed analysis of the "compatibility"[17] of such purposes). Similarly, a researcher that collected data for one study might be precluded from using the same data for another study. Because establishing another legal basis may be impractical or impossible, this exemption provides researchers with greater flexibility to use personal data for research purposes where research may not have been the primary reason for the data's collection.

Second, the research exemptions lift restrictions on using sensitive personal data for research purposes, provided there is a basis in EU or Member State law. Processing sensitive personal data under the GDPR is generally forbidden, except under narrowly circumscribed conditions, such as where a data subject has provided "explicit consent" or where personal data is "manifestly made public."[18] In addition to these conditions, the GDPR permits a controller to process sensitive personal data where "processing is necessary for [qualifying research] in accordance with Article 89(1)."[19] This is of critical importance for many forms of research that rely on health data, information concerning race, ethnicity, sexual orientation or religion, or research concerning political beliefs. However, in order to make use of this condition, the research must be "based on Union or Member State law which shall be proportionate to the aim

[15] GDPR, Art. 5(1)(b).

[16] *Id.*

[17] GDPR, Art. 6(4).

[18] GDPR, Art. 9(1).

[19] GDPR, Art. 9(1)(j).

pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject."[20]

Third, the research exemptions relax the application of certain individual rights in the research context and permit EU or Member State law to introduce additional derogations.[21] In particular, controllers may override requests to erase personal data where applying the right to erasure "is likely to render impossible or seriously impair the achievement of the [qualifying research] objectives . . . ."[22] Similarly, where a researcher obtains personal data from a source other than from the data subject directly, the obligation to inform the data subject about the intended processing would not apply if it would "render impossible or seriously impair the achievement of the [qualifying research] objectives . . . ."[23] Other rights, such as the right to object and rights of access, rectification, and restriction may also be limited in the context of qualifying research, but only where other EU or Member State laws so provide.[24]

In order to rely on the research exemptions, controllers that process personal data for qualifying research purposes must implement "appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject."[25] What precisely those safeguards must be is not specified, but they must include "technical and organisational measures . . . in particular in order to ensure respect for the principle of data minimisation [and] may include pseudonymization provided that [the qualifying research] purposes can be fulfilled in that manner."[26] In addition, the GDPR requires controllers to use pseudonymized or anonymized data (such as de-identified or aggregated data sets) "[w]here [the qualifying research] purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects."[27] A GDPR recital further states that research should be conducted "in keeping with recognised ethical standards for scientific research."[28]

---

[20] *Id.*

[21] GDPR, Art. 17(3)(d), 89(2).

[22] GDPR, Art. 17(3)(d).

[23] GDPR, Art. 14(5)(b).

[24] GDPR, Art. 89(2).

[25] GDPR, Art. 89(1).

[26] *Id.*

[27] *Id.*

[28] GDPR, Recital 33.

3. Using data for independent research

The GDPR's research exemptions were intended to strike a balance between, on the one hand, facilitating research, and on the other, protecting the rights to data protection of affected individuals. Although the research exemptions provide meaningful support for research activities, in some contexts, research activities have been hampered or slowed by insufficient guidance, particularly regarding predicate concepts such as anonymization, the presumption of compatibility, and the definition of qualifying research.[29] In addition, the fact that some of the research exemptions require additional EU or Member State legislation has also reduced the utility of such exemptions where the necessary authorizing legislation is inconsistent across the EU or, more often, nonexistent. For these reasons, legal scholars have concluded that "the impact of the exemptions is likely to be limited in practice."[30]

The challenges for public and private sector organizations that aim to supply data for independent research, ("**data originators**"), including digital services such as Facebook, are even more acute. This is because the GDPR often contemplates a holistic assessment of the lifecycle of data processing activities that is difficult to achieve from the vantage point of a data originator. The GDPR's accountability principle, which places the burden on controllers to *demonstrate* compliance with the Regulation, arguably imposes obligations on data originators to ensure that privacy is adequately protected when sharing data with third parties.

As explored in detail below, key provisions of the GDPR create tension between the actions data originators are encouraged to take to minimize data protection risks and the independence of researchers. In particular, as explained in Section III.1.a. below, identifying an appropriate legal basis for data sharing requires the involvement of

---

[29] For example, the application of GDPR provisions to clinical trials in the EU has been the subject of significant debate, particularly in light of overlapping requirements of the Clinical Trials Regulation. To add complexity, the lack of harmonization at the Member State level on the conditions for permitting clinical trials has led to fragmentation and challenges for cross-border studies. *See* European Data Protection Board, *Opinion 3/2019 concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection regulation (GDPR) (art. 70.1.b))* , at 4-7 (Jan. 23, 2019).

[30] Miranda Mourby et al., *Governance of academic research data under the GDPR—lessons from the UK*, 9 International Data Privacy Law, 192, 201 (Aug. 2019).

the data originator. For example, where the legal basis is consent, the data originator needs to obtain consent with sufficient specificity to meet the high standards EU regulators have articulated. Where the legal basis is "legitimate interests," the data originator must be satisfied that the subsequent processing will not result in disproportionate risks to data subjects. Transparency requirements, described in Section III.1.b., raise obstacles over whether data originators should be able to adequately explain the resulting processing by independent researchers. And, accountability requirements—such as the requirement to implement risk- and context-appropriate safeguards described in Sections III.1.c. and III.2.b.—encourage data originators to impose contractual limitations and to supervise the data processing activities of researchers. There are challenges over whether data originators must even exercise oversight over the release of highly de-identified datasets, where they cannot be certain that the technical processes applied to the data meet the nebulous and inconsistent anonymization standards EU regulators have articulated, as described in Section III.4.

While it may best serve users' privacy interests for data originators to retain tight control over personal data, maintaining the independence of researchers is critical to enable the forms of beneficial research the European Commission has sought to promote. This is particularly true where research concerns the effects of digital services themselves on issues of public concern, such as elections and democracy. It should be up to independent experts—and not the digital services that build the tools—to determine what issues should be studied.[31]

These challenges do not preclude data originators from sharing data for independent research altogether. Rather, they discourage data originators from doing so because of the need to develop and implement technical and governance safeguards—often at significant cost to the data originator—in the absence of clear and specific guidance for appropriately addressing data protection risks. As a result, even data originators determined to promote research may be inclined to limit the availability of scientifically valuable information or impose seemingly excessive restrictions on access. While these efforts are important to data protection, they do restrict the provision of data to independent researchers and may require data originators to

[31] European Commission, *A European strategy for data*, at 14 COM (2020) 66 final (Feb. 19, 2020)

exercise oversight in ways that could threaten the independence of some research activities. Moreover, even after investing significant resources to implement such measures, data originators may nonetheless face some risk due to the uncertainty of how regulators and courts might interpret the relevant provisions. As a result, many organizations that have scientifically valuable information choose not to share it with independent researchers.

Below, we consider the primary drivers of legal challenges for data originators before exploring the ways in which we believe EDMO's proposed code of conduct could alleviate some of these challenges.

III. DATA SHARING CHALLENGES UNDER THE GDPR

This section first considers the challenges that undermine the provision of data by data originators to independent researchers. Next, we consider the application of the research exemptions and explain why these exemptions fail to provide data originators with sufficient legal certainty to overcome GDPR risks. We also consider how these challenges are compounded by a lack of harmonization among Member State laws and the absence of clear standards for the anonymization of personal data.

1. The GDPR's analytical framework discourages data originators from sharing personal data for independent purposes.

When processing personal data, including to make such data available for independent research, a controller must comply with, and is responsible for demonstrating compliance with, the GDPR's core principles set out in Article 5.[32] These core principles require controllers to inform individuals of the nature of the processing activity, process personal data only for those informed and lawful purposes, limit the collection of personal data to the extent necessary to achieve the intended purposes, store personal data for only as long as necessary, protect the security of personal data, and at all points, take responsibility for the processing of personal data and for demonstrating compliance with the foregoing. Here, we focus on three categories of requirements that pose particular challenges for sharing data that includes personal data for independent research: (a) establishing a legal basis for processing, (b) transparency, and (c) accountability.

a. The GDPR's legal bases for processing do not easily align with sharing data for research purposes.

To process personal data for any purpose, a controller must identify an appropriate "legal basis" that permits the processing.[33] As the term "processing" is defined

[32] GDPR, Art. 5(2).

[33] GDPR, Art. 6(1).

broadly to include "any operation or set of operations which is performed on personal data," the "disclosure by transmission, dissemination or otherwise making available" of personal data to an independent party requires the disclosing party to identify a legal basis for the disclosure.[34] Two legal bases are most likely to apply to the disclosure of personal data for research, but each presents its own set of unique challenges for data provided: consent and legitimate interests.

Consent may not be possible or appropriate for all forms of socially beneficial research.

The GDPR permits controllers to process personal data where the relevant data subject has given consent to the processing. Consent plays an important role in data protection because it grants data subjects the ability to control the processing of their personal data, which helps to promote trust and legitimacy.[35] Provided consent is obtained lawfully, a data originator would gain greater certainty that it could share personal data as authorized by the data subject, including for research purposes. This makes consent particularly valuable for researchers.[36]

However, obtaining consent also poses fundamental challenges for some forms of research. Consent under the GDPR must be specific, informed, freely-given, revocable, and granted by an unambiguous affirmative action.[37] To meet the GDPR's specificity requirement, consent for "different personal data processing operations" should be given separately.[38] As explained by the European Data Protection Board ("**EDPB**"), this means that "specific consent can only be obtained when data subjects are specifically informed about the intended purposes of data use concerning them."[39] But the EDPB's strict construction of this requirement extends further by closely

---

[34] GDPR, Art. 4(2).

[35] Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 Nw J. Tech & IP 239 (2013).

[36] *See, e.g.,* U.S. Dep't of Health & Human Servs., Federal Policy for the Protection of Human Subjects.

[37] GDPR, Art. 7.

[38] GDPR, Recital 43.

[39] EDPB, *Guidelines 05/2020 on consent under Regulation 2016/679* of 4 May 2020, 14.

linking "specificity" to the concept of control over each activity.[40] As a result, "consent can only be an appropriate lawful basis if a data subject is offered control and is offered a genuine choice with regard to accepting or declining the terms offered or declining them without detriment."[41] Not only could this strict construction result in a requirement to obtain consent for each specific research protocol (rather than permitting data subjects to consent to social scientific research generally), but granting individuals this level of control could also skew research results, particularly in the big data research context, as the sample of individuals who opt-in to any particular study may not be representative of the population as a whole.[42]

Recognizing that "[i]t is often not possible to fully identify the purposes of personal data processing for scientific research purposes at the time of data collection," the GDPR permits data subjects to "give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research."[43] But how specifically those "areas of scientific research" would need to be described is unclear. On the one hand, EDPB guidance suggests it would be sufficient for data subjects to understand "the state of play," yet the same guidance states that "having a comprehensive research plan available for data subjects to take note of, before they consent could help to compensate [for] a lack of purpose specification."[44] Providing this level of specificity at scale may not be practicable for data originators. The research exemptions provide some relief, but, as discussed in Section III.2 below, they may not be sufficient to overcome the challenges identified here.

---

[40] *Id.* at 14 ("Notwithstanding the provisions on compatibility of purposes, consent must be specific to the purpose. Data subjects will give their consent with the understanding that they are in control and their data will only be processed for those specified purposes. If a controller processes data based on consent and wishes to process the data for another purpose, too, that controller needs to seek additional consent for this other purpose unless there is another lawful basis, which better reflects the situation.").

[41] *Id.* at, 5.

[42] Joseph W. Sakshaug et al., *Evaluating Active (Opt-In) and Passive (Opt-Out) Consent Bias in the Transfer of Federal Contact Data to a Third-Party Survey Agency*, 4:3 J. of Survey Stat. and Methodology (Sept. 4, 2016), https://academic.oup.com/jssam/article-abstract/4/3/382/2399768 (concluding that the opt-out consent does a better job of minimizing self-selection bias and maximizing the validity of the survey estimates compared with active consent procedures, but that both consent procedures increase the total self-selection bias).

[43] GDPR, Recital 33.

[44] EDPB, *Guidelines 05/2020 on consent under Regulation 2016/679*, at 31 (May 4, 2020).

A further challenge with relying on consent for qualifying research is that the GDPR grants data subjects a right to withdraw consent at any time.[45] As noted by the EDPB, the "withdrawal of consent could undermine types of scientific research that require data that can be linked to individuals, however the GDPR is clear that consent can be withdrawn and controllers must act upon this—there is no exemption to this requirement for scientific research."[46] In practice, however, researchers often need to retain the data underlying any study to validate and evidence the legitimacy of their results internally or through external validators (*e.g.*, scientific journals prior to publishing the scientific report study).[47] Data retention could also be justified to enable further developments of the initial research. Depending on the research, there can be statutory retention periods of the data that evidence the legitimacy of the scientific results. None of these circumstances are taken into account by the EDPB, which concluded that "[i]f a controller receives a withdrawal request, it must in principle delete the personal data straight away if it wishes to continue to use the data for the purposes of the research."[48]

Although consent may be an appropriate basis for some research, these requirements pose special challenges for data originators, both in relation to the independence of researchers' activities, and the scale of data originators' intended data sharing. Even if formal consent requirements could be met in relation to any one individual's personal data, where research involves interactions between individuals, obtaining consent from all individuals whose personal data could be implicated may prove impossible. A study of the propagation of disinformation on social media, for example, may be seriously impaired if data originators or researchers were required to obtain

---

[45] GDPR, Art. 7(3).

[46] EDPB, *Guidelines 05/2020 on consent under Regulation 2016/679*, at 32 (May 4, 2020).

[47] In addition, deletion in the research context could, paradoxically, weaken privacy protections in de-identified data sets. Not only could smaller sample sizes result in data sets that are easier to reidentify—for example, where researchers study smaller scale phenomena—but any observed changes to the output resulting from removing one or more individuals' data could reveal something about those individuals as well as the data subjects remaining within the study. Deletion may be especially problematic for data sets that have been de-identified using techniques that involve adding noise, such as differential privacy, since this may reveal the added noise and undermine the intended protections.

[48] *Id.* It is not only the published results that are of importance to researchers. In fact, to ensure that any study is replicable and to maintain the integrity of social scientific research, researchers may need to maintain data that is not always possible to fully anonymize. Adhering to this strict deletion right, therefore, could impair social scientific research even if the published results fall outside the scope of the right.

consent from anyone who interacted with specific content. Two challenges are important to consider. First, it may be impractical to consent an entire network of people in order to study knock-on effects of disinformation on networks. Second, the very people who are of most interest to researchers—those with malintent—are also given a direct pathway to remove scrutiny of their activities: not providing consent.

Legitimate interests requires an assessment that could compromise research independence.

The GDPR also permits the processing of personal data, without consent, where it is "necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject . . . ."[49] To qualify for processing under the "legitimate interests" legal basis, a controller must satisfy a three-part test: first, the controller must demonstrate that interest being pursued is legitimate; second, the controller must demonstrate that the proposed processing of personal data is necessary to pursue the legitimate interest; and finally, those interests must be balanced against any risks to the rights and freedoms of concerned individuals.[50]

Although, in theory, legitimate interests may be an appropriate basis for research, in practice, this basis poses special challenges in the context of data sharing for independent research. First, to rely on this basis, data protection authorities have stated that the interest being pursued must be "sufficiently clearly articulated to allow the balancing test to be carried out against the interests and fundamental rights of the data subject."[51] In addition, the data shared for such purposes must be limited to what's "necessary." Taken together, these elements impose obligations on data originators that are challenging to meet without exerting some level of control over research activities.

The balancing exercise that forms the third element of the legitimate interests test also poses challenges because it requires controllers to conduct a holistic assessment of the particular risks to individuals that an activity poses, in light of the processing purposes, the safeguards that will be put in place, and the reasonable expectations

[49] GDPR, Art. 6(1)(f).

[50] Case C-13/16, Valsts policijas Rīgas reģiona pārvaldes Kārtības policijas pārvalde v Rīgas pašvaldības SIA 'Rīgas satiksme', ECLI:EU:C:2017:336 ¶ 28 (May 4, 2017).

[51] Article 29 Data Protection Working Party *Opinion No. 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC*, at 24 (Apr. 9, 2014).

of such affected individuals.[52] In the context of research concerning the Covid-19 outbreak, for example, data protection authorities have specified that a lawfulness analysis should include, among other elements, "specifying the research questions and assessing the type and amount of data necessary to properly answer these research questions."[53] Data originators (who are themselves likely not scientists) are unlikely to be in a position to undertake this analysis in a rigorous manner or without compromising the independence of the research.

EU data protection authorities have acknowledged the difficult position of entities that make personal data available to third parties without strict control of its subsequent use, but have nonetheless endorsed interpretations that make data sharing for beneficial purposes more challenging. For example, in an exchange of letters with the Internet Corporation for Assigned Names and Numbers ("**ICANN**") concerning the publication of names and contact details for domain registrants (*i.e.* WHOIS Data), the EDPB argued that ICANN's specifications did not satisfy the legitimate interests test because, once the data was made available, it could be used by third parties in unknown ways.[54] The EDPB's position was especially striking because it recognized the primary purposes for the publication of WHOIS Data—namely, to facilitate the investigation of criminal conduct and to permit third-parties to pursue intellectual property violations—were important societal interests and the data categories within Whois Data were limited only to non-intrusive fields, such as names and contact details.

Efforts by data originators to address these requirements have led to resistance from the research community. For example, as part of a Facebook partnership with Social Science One, a body set up by several nonprofit foundations to facilitate research on the impact of social media on democracy and elections, Facebook provided access to an aggregated database of URLs that had been shared by Facebook users.[55] In order to access the database, researchers were required to submit their research proposals to Social Science One, which in turn would decide which proposals would be granted access. Researchers that were permitted to access Facebook data had to agree not to use Facebook data except for the permitted research purpose, unless otherwise

---

[52] GDPR, Recital 47.

[53] EDPB, *Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak*, at 10 (Apr. 21, 2020).

[54] *See* EDPB, Letter to ICANN (Jul. 5, 2018).

[55] Social Science One, *Social Science One: Public Launch*, Harvard Univ. (Jul. 11, 2018).

authorized by Facebook.[56] These measures, carefully designed to protect user privacy, were nevertheless criticized for undermining researchers' independence, slowing down approval processes, centralizing control of data with a few (American) institutions, and limiting the value of the shared data itself.[57] This experience provided us with an apt example of how difficult it can be to balance privacy and research, even with the best of intentions.

At the same time, the legal certainty provided by a code of conduct, and the opportunity to develop standards based on input from all relevant stakeholders, could help address these challenges within the parameters of the legitimate interests balancing test. For instance, the resulting code of conduct could permit data originators to release data that is de-identified to specified standards more widely, and with lower levels of oversight and control. By contrast, more detailed vetting and oversight may be required before researchers are granted access to other data sets. In addition, the code of conduct could establish an independent body responsible for making these assessments. This could allow data originators to satisfy their obligations without themselves determining when a researcher's access should be granted and under what conditions.

Other legal bases have limited utility outside narrow use cases.

Other legal bases provided by the GDPR could apply to some limited research activities. For example, where health and safety is at stake, research may also be permitted where "necessary in order to protect the vital interests of the data subject or of another natural person."[58] Researchers may also be able to rely on the "public interest" legal basis where there is an underlying legal mandate in EU or Member State law that supports a particular study.[59] Some have suggested that research affiliated with a public research institution may qualify, but this would be subject to a similar necessity and proportionality analysis as required to rely on legitimate interests.[60]

---

[56] Social Science One, Draft Research Data Agreement.

[57] Axel Bruns, *After the 'APIcalypse': social media platforms and their fight against critical scholarly research*, 22 Info., Commc'n & Soc'y 11 (Jul. 11, 2019), https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1637447.

[58] GDPR, Art. 6(1)(d).

[59] GDPR, Art. 6(1)(e).

[60] Miranda Mourby et al., *Governance of academic research data under the GDPR—lessons from the UK*, 9 Int'l Data Priv. L., 192 (Aug. 2019).

Similarly, the "legal obligation" legal basis could apply in some circumstances if an EU or Member State law requires certain research to be undertaken. However, even where such bases could apply to a researcher, there may be limitations on a data originator's ability to make use of such bases if Member State laws are inconsistent and no EU-level mandate is provided.

b. *Although researchers may benefit from more relaxed transparency requirements, it is not clear that data originators can also take advantage of these provisions.*

The GDPR requires controllers to provide disclosures concerning the nature of any data processing operations, including clear information about the purposes of processing and the categories of recipients of any personal data collected.[61] Where personal data is collected from the data subject directly, such disclosures must be provided "at the time when personal data are obtained."[62] To meet transparency requirements, the disclosures must be sufficiently specific such that "the data subject should be able to determine in advance what the scope and consequences of the processing entails and that they should not be taken by surprise at a later point about the ways in which their personal data has been used."[63] Moreover, any incompatible changes to the purposes of personal data processing after notice has been provided must be communicated to data subjects.[64]

Similar to the questions raised by consent requirements as described above, it may be challenging for data originators to meet transparency requirements with respect to, on the one hand, the existence of future potential researchers unknown at the time of the data collection and, on the other hand, subsequent use of personal data by researchers without limiting their independence. For example, in its guidance on transparency, the EDPB explained that, although the GDPR text permits a controller to describe the "categories of recipients" of personal data, controllers must provide more "meaningful" information on data sharing to comply with "the principle of fairness":

---

[61] GDPR, Arts. 13, 14.

[62] GDPR, Art. 13(1).

[63] Article 29 Data Protection Working Party, *Guidelines on transparency under Regulation 2016/679*, at 7 (Apr. 11, 2018).

[64] *Id.* at 16-17.

> "In practice, this will generally be the named recipients, so that data subjects know exactly who has their personal data. If controllers opt to provide the categories of recipients, the information should be as specific as possible by indicating the type of recipient (i.e. by reference to the activities it carries out), the industry, sector and sub-sector and the location of the recipients."[65]

Providing this level of detail could present practical challenges for data originators if they grant a broad range of researchers access to data, so it will be important for EDMO and the relevant stakeholders to consider how to balance the preference for specificity against the desire for broadly available research data. The more widely the data is made available, the more the challenges of providing appropriate transparency are exacerbated. For data originators, this creates an incentive to limit the extent of data sharing for research purposes, which, paradoxically, reduces the transparency of the overall system of data processing, as fewer researchers are able to access and interrogate the data.

Even when the underlying data itself is not made openly available, the push toward Open Science, including as encouraged by the European Commission,[66] often requires researchers to publish their data along with their findings or make data available to others who may seek to replicate the research, which could expose such personal data to an unlimited number of recipients. It is not clear that general notice that personal data will be made publicly available would be sufficient to address such requirements.[67] Some EU regulators have called into question whether the GDPR's transparency requirements can be met where personal data is either made openly available or shared with a range of independent parties.[68]

The GDPR provides an exception to transparency requirements within the context of social scientific research where providing information to data subjects would prove

---

[65] *Id.* at 37.

[66] European Commission, Open Science, last accessed Nov. 13, 2020, https://ec.europa.eu/research/openscience/index.cfm.

[67] See Article 29 Data Protection Working Party, *Opinion 3/2013 on purpose limitation*, at 18-19 (April 2, 2013); Article 29 Data Protection Working Party, *Opinion 06/2013 on open data and public sector information ('PSI') reuse*, at 9 (June 5, 2013).

[68] *See,e.g.,* United Kingdom Information Commissioner's Office, Update report into adtech and real time bidding, 2019 at 19.

impossible or require disproportionate effort.[69] Although this exception is expressly designed to facilitate research, it is of limited utility to data originators that collect personal data directly from data subjects.[70] The exception applies only where data is obtained indirectly, from third-party sources, *i.e.*, for the researchers. Therefore, some researchers may be able to rely on the exception when they receive personal data from a data originator, but a data originator that collects personal data directly cannot.[71] Another exception that applies where "disclosure is expressly laid down by [EU] or Member State law," could help address these challenges, but to date, no EU-level law provides a generally-applicable basis for exempting the disclosure of platform data for research purposes from transparency requirements.[72]

c. Data originators are not well-positioned to meet the GDPR's accountability requirements where they cannot analyze the risks of data use by independent parties.

---

[69] GDPR, Art. 14(5)(b).

[70] Article 29 Data Protection Working Party, *Guidelines on transparency under Regulation 2016/679*, at 30 (April 11, 2018)("Given the emphasis in Recital 62 and Article 14.5(b) on archiving, research and statistical purposes with regard to the application of this exemption, WP29's position is that this exception should not be routinely relied upon by data controllers who are not processing personal data for the purposes of archiving in the public interest, for scientific or historical research purposes or statistical purposes.").

[71] *Id.* ("The only difference between an Article 13 and an Article 14 situation is that in the latter, the personal data is not collected from the data subject. It therefore follows that impossibility or disproportionate effort typically arises by virtue of circumstances which do not apply if the personal data is collected from the data subject. In other words, the impossibility or disproportionate effort must be directly connected to the fact that the personal data was obtained other than from the data subject.").

[72] GDPR, Art. 14(5)(c); *see also* EDPB, *Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak*, at 9 (Apr. 21, 2020) ("Article 14 (5) (c) GDPR allows for a derogation of the information requirements in Articles 14 (1), (2) and (4) insofar as the obtaining or disclosure of personal data 'is expressly laid down by Union or Member State law to which the controller is subject'. This exemption is conditional upon the law in question providing 'appropriate measures to protect the data subject's legitimate interests'. As stated in the above mentioned Transparency Guidelines, such law must directly address the data controller and the obtaining or disclosure in question should be mandatory upon the data controller.").

In a significant departure from the requirements of the Directive, the GDPR requires organizations to demonstrate compliance with all requirements and to implement appropriate safeguards that are tuned to the risk inherent in any processing activity. At root, the GDPR's "risk-based approach" requires controllers to assess the lawfulness of a processing activity by taking into account the "nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons."[73] The relevant risks for this analysis include "discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data, unauthorised reversal of pseudonymisation, or any other significant economic or social disadvantage."[74]

To comply with the GDPR's accountability requirements, controllers must implement appropriate controls, taking the context into account, to protect the security of personal data and prevent the processing of personal data in breach of GDPR requirements.[75] Controls could include technical measures, such as security standards and privacy-enhancing techniques (e.g. encryption or pseudonymization), as well as organizational controls, such as access limitations, use limitations, and mechanisms for permitting data subjects to choose how their personal data will be used and shared, which are difficult to apply in the context of independent research.[76]

These challenges are exacerbated by new and untested provisions of the GDPR's accountability regime that might be read to unduly extend liability to such data sharing. For instance, the EDPB's guidance on data protection suggests that EU data protection authorities might consider there to be "an obligation on the original controller not to make the personal data unduly accessible in the first place."[77]

---

[73] GDPR, Art. 24.

[74] GDPR, Recital 75.

[75] GDPR, Art. 32.

[76] While these controls must expressly extend to a controller's selection of a processor, some interpretations of the GDPR leave open the extent to which a controller remains responsible for the subsequent processing of personal data by a recipient acting as a controller. GDPR, Arts. 19 and 26.

[77] EDPB, *Guidelines 4/2019 on Article 25 Data Protection by Design and by Default*, at 13 (Nov. 13, 2019)

2. The GDPR's research exemptions fail to provide sufficient legal assurances for data originators.

Section III.1 highlighted several of the GDPR requirements that create tension and legal challenges for data originators that supply personal data for independent research. This section turns to the GDPR's research exemptions and explains why they fail to provide sufficient legal certainty to permit data originators to share personal data for independent research without oversight and technical safeguards.

a. What type of research is qualifying research under the GDPR is unclear.

As a threshold matter, in order for the GDPR's research exemptions to apply, the research in question must qualify as "scientific research," "historical research" or "statistical purposes" under the GDPR. The GDPR does not specifically define "research," but it states that the concept of research should be interpreted "in a broad manner" and would encompass "technological development and demonstration, fundamental research, applied research and privately funded research."[78] In addition, the GDPR specifies that the concept of research should take into account the objective stated in the Treaty on the Functioning of the European Union "of achieving a European Research Area."[79] As the purpose of the European Research Area is to enable "researchers, scientific knowledge and technology circulate freely, and encouraging it to become more competitive, including in its industry, while promoting all the research activities deemed necessary by virtue of other Chapters of the Treaties," the concept of research could apply to a broad range of subject matter.[80]

A preliminary opinion from the European Data Protection Supervisor ("**EDPS**") emphasizes that one of the criteria which determines the application of the research exemptions is adherence to "the relevant sectoral standards of methodology and

[78] GDPR, Recital 159.

[79] *Id.*

[80] TFEU Treaty, Art. 179 .

ethics,"[81] citing guidance from the former Article 29 Working Party (the precursor body to the current EDPB), which explains scientific research as "a research project set up in accordance with relevant sector-related methodological and ethical standards."[82] Because not all research will be qualifying research, data originators may face legal risk if they share data in reliance on the research exemptions without first vetting that a research proposal qualifies. This result could undermine the independence of the resulting research. An independent process for determining when the research exemptions apply may be necessary to give effect to these provisions without impairing broad, blue sky work that may not be politically popular at any given moment. In any event, legal certainty as to the definition of scientific research is indispensable.[83]

b. *The purpose limitation exemption for qualifying research requires a careful assessment of appropriate safeguards.*

Although the research exemptions may provide meaningful flexibility for researchers, they do not provide relief from the core tensions and challenges described in Section III.1. The primary benefit for researchers is a relaxation of the GDPR's "purpose limitation" rule, which ordinarily limits the use of personal data to "specified, explicit and legitimate purposes" and prevents the further use of such data "in a manner that is incompatible with those purposes."[84] Since qualifying research is not "considered

---

[81] European Data Protection Supervisor, *A Preliminary Opinion on data protection and scientific research*, at 12 (Jan. 6, 2020). The full list of criteria is: 1. Personal data is being processed, 2. Relevant sectoral standards of methodology and ethics apply, including the notion of informed consent, accountability and oversight; and 3. The research is carried out with the aim of growing society's collective knowledge and wellbeing, as opposed to serving primarily one or several private interests.

[82] See the Article 29 Data Protection Working Party, *Guidelines on consent under the GDPR* 2016/679, at 27 – 30 (Apr. 10, 2018).

[83] In this regard, in contrast to the narrow concept of research articulated by data protection authorities, EU law appears to permit a broader construction. For example, Council Directive 2005/71/EC, art. 2, 2005 O.J. (L 289) 17 (EU), defines scientific research broadly to include "creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications." Given that qualifying research is not precisely defined within the GDPR and is explained by reference to other principles of EU law, this broader definition should be taken into account.

[84] GDPR, Art. 5(1)(b).

to be incompatible with the initial purposes," this permits controllers to use personal data collected for a one set of purposes—including commercial purposes—to subsequently be used for research.[85]

The potential benefits of this exemption are apparent from the discussion of GDPR's legal bases for research in Section III.1a. Where personal data is collected for one purpose, the GDPR permits such data to be processed for another, secondary purpose, if the secondary purpose is "compatible" or (satisfies another element of Article 6 GDPR).[86] Compatibility is assessed on the basis of a series of factors—including the context, the link between the purposes, the possible consequences for individuals, and the safeguards in place—that bear similarity to the balancing test for relying on legitimate interests.[87] If, however, an activity is deemed to be compatible, "no legal basis separate from that which allowed the collection of the personal data is required."[88] The research exemption allows controllers to presume that qualifying research is compatible.[89]

To benefit from the relaxed purpose limitation principle under the research exemptions, qualifying research must be subject to "appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject."[90] As explained by the EDPS:

> "[t]he presumption [of compatibility] is not a general authorization to further process data in all cases for historical, statistical or scientific purposes. Each case must be considered on its own merits and circumstances. But in principle personal data collected in the commercial context, for example, may be further used for scientific research purposes, by the original or a new controller, if appropriate safeguards are in place."[91]

---

[85] *Id.*

[86] GDPR, Art. 6(4).

[87] *Id.*

[88] *Id.* at Recital 50.

[89] European Data Protection Supervisor, *A Preliminary Opinion on data protection and scientific research*, at 22 (Jan. 6, 2020).

[90] GDPR, Art. 89(1).

[91] European Data Protection Supervisor, *A Preliminary Opinion on data protection and scientific research*, at 22 (Jan. 6, 2020).

The absence of clear standards for what types of safeguards will be considered appropriate *ex ante* complicates the application of this exemption for data originators. Not only does this require data originators to conduct a detailed and fact-specific analysis, the GDPR also states that a controller relying on the research exemptions must consider pseudonymizing or otherwise de-identifying the relevant data "[w]here [the research] purposes can be fulfilled . . . in that manner."[92] Thus, to conduct this analysis, a data originator may need to take the researcher's study design into account to understand what data is necessary to share for a given study, a result which could cause friction for some researchers.[93]

3. Variation in Member State law exacerbates risks for cross-border research.

In addition to the challenges data originators face in implementing appropriate and targeted safeguards where they do not control the research protocols of independent researchers, variation in Member State law adds further complexity where independent researchers may be in multiple Member States. This complexity is heightened for research involving special categories of personal data, as the use of such data for research purposes often depends on specific Member State laws.

a. Key research provisions are not harmonized across the EU.

Although the GDPR permits controllers to process sensitive personal data for research purposes, this is only permitted where the research is "based on Union

---

[92] Id.

[93] Axel Bruns, *Facebook shutes the gate after the horse has bolted, and hurts real research in the process*, Internet Pol'y Rev. (Apr. 25, 2018) (arguing that "[t]he narrow terms of reference for this initiative (elections and democracy), the requirement to adhere to a research agenda defined by the selection panel, and the selection process itself are inherently excluding a much broader range of research that investigates the impact of Facebook on all aspects of society."). *See also* Cornelius Puschmann, *An end to the wild west of social media research: a response to Axel Bruns*, 22:11 Info., Commc'n & Soc'y, 1588, 1589 (2019) ("[T]here is currently no suitable alternative model in place able to provide platform data to academics in a fashion that both ensures high standards of representativeness and reproducibility, and at the same time respects user privacy. . . . [C]ommercial API data is biased, incomplete, and subject to a range of awkward technical and contractual restrictions that impede its usefulness for empirical research. If anything, new mechanisms for controlled access to data for research purposes could in the future serve as a model for developer APIs, rather than the other way around. Furthermore, open APIs that anyone can readily use will in many cases prove incompatible with stringent legal requirements for privacy and data security.").

or Member State law."[94] In the absence of a broad legal mandate at the EU level that generally authorizes research involving sensitive personal data, reliance on this provision often depends on Member State implementations of the GDPR that may authorize particular research topics.[95]

Some of the most critical forms of research involve the processing of sensitive personal data.[96] For instance, research into the dissemination of political opinions or the impact of media on elections and democracy, may require the use of data that is subject to these heightened restrictions under the GDPR.

And yet, the rules for processing sensitive personal data for research purposes vary considerably across EU Member States.[97] In the Netherlands, for example, any processing of sensitive personal data for research purposes requires consent, unless "asking for express consent proves impossible or requires disproportionate effort."[98] By contrast, the Danish Data Protection Act does not require express consent to process sensitive data for research purposes if the processing of the sensitive data is significantly relevant for the public.[99] Ireland has adopted a different approach from the two: consent is provided as an example of a "suitable and specific measure," but is not strictly required provided other measures are in place, such as encryption and specific training for those handling personal data.[100] Where research involves researchers throughout multiple Member States, applying the appropriate standards to such research could lead to conflicting results.

It is not only the provisions that apply to sensitive personal data that may depend on Member State law. Leaving aside Art. 14.5(b) of GDPR (information when data are not directly collected from data subjects) and 17.3(d) (erasure), research exemptions that exclude the application of data subject rights to qualifying research, such as rights to

---

[94] GDPR, Art. 9(2)(j).

[95] Edward Dove, *The EU General Data Protection Regulation: Implications for International Scientific Research in the Digital Era*, 46(4) J L Med & Ethics 1013–30 (2018).

[96] *Id.*

[97] Edward S. Dove and Jiahong Chen, *Should consent for data processing be privileged in health research? A comparative legal analysis*, Int'l Data Priv. L. (Feb. 25, 2020).

[98] See Uitvoeringswet Algemene verordening gegevensbescherming, Art. 24 (May 25, 2018) (translated from Dutch to English).

[99] Databeskyttelsesloven, No. 502 of 23/05/2018, ¶ 10(1) (translated from Danish to English).

[100] Irish Data Protection Act 2018 (No. 7/2018) § 36.

access, rectification, restriction and to object to processing, also depend on Member State implementations of the GDPR.[101] This means that most individuals' rights in relation to their personal data may vary depending on which Member State law applies. While data originators could seek a consistent approach by applying the highest standards across the board, as the examples above illustrate, the standards applied in different Member States may not always be higher or lower—sometimes they are just different.

b. *Overlapping Member State implementations of the GDPR and other local legal frameworks could apply to cross-border research.*

An additional challenge for cross-border research is to determine which Member State's laws apply. Curiously, while the GDPR specifies when a Member State's regulators would have authority to enforce and in which Member States a data subject could bring a claim, the GDPR does not state explicitly when a Member State's laws would apply to processing. Some Member States have adopted their own rules for determining which Member State laws to apply, and in some cases, the rules of varying Member States overlap and conflict. This poses foundational challenges for some forms of cross-border research because not only do many of the GDPR provisions that govern research depend on Member State law but also local laws unrelated to GDPR might be those that actually regulate a specific kind of research.[102]

For example, Spain's implementation of the GDPR applies only where an organization is either processing personal data in the context of the activities of an establishment in Spain or where an organization with no EU establishment offers goods or services to, or monitors the behavior of, individuals in Spain.[103] This means that Spanish law may not apply to an organization established in another EU Member State but not in Spain. By contrast, the German Federal Data Protection Act applies in comparable situations as well as whenever personal data is physically processed in Germany. Therefore, an entity established in Spain could be subject to both Spanish and German

[101] *See* GDPR, Art. 89(2).

[102] One example of this is in the context of health research where Member State medical confidentiality laws, which are not harmonized at the EU level, often overlap and intersect with data protection requirements.

[103] Law on the Protection of Personal Data and the Guarantee of Digital Rights (B.O.E, 2018, 16673) (Spain).

law if it uses technical infrastructure in Germany. Other Member States' implementing

legislation, such as Ireland, are silent on when national law would apply, which furthers ambiguity as to when those Member States' national provisions and derogations would apply.

The range of applicable law standards has important implications for provisions of the GDPR that depend on national implementing legislation, such as the conditions for processing special categories of personal data. In the context of big data research, which—involves data originators and research institutions in multiple locations, including outside the EU—it may not be possible to know the most appropriate method to tackle compliance. This complicates any attempts to segregate research activities that may be lawful in one Member State but not in another. To limit risks, data originators may need to apply risk mitigation strategies, such as providing access to only limited or aggregated information, or restricting researchers from using data for certain purposes that could conflict with Member State requirements. This places a burden on independent researchers as to whether they are permitted to receive certain data sets, in accordance with their local Member State laws. As such, a data originator may be reluctant to authorize certain studies—even where the topic of research is expressly authorized by one Member State—due to the risks that the more restrictive laws of another Member State could apply to independent researchers.

4. The standards for effective anonymization under the GDPR remain unsettled.

The GDPR encourages the anonymization of personal data, not only as a way controllers can mitigate legal and privacy risks, but also as a safeguard controllers could consider when relying on some of the research exemptions. However, the standards of anonymization are subject to ambiguous standards, which undermine data originators' ability, in certain circumstances, to determine with certainty that data has been fully anonymized.

a. Guidance from EU regulators adopts strict and inconsistent standards of anonymization.

The GDPR does not explicitly define anonymization, but the concept of anonymization emanates from the inverse of the GDPR's definition of personal data. In deciding whether an individual is identifiable from information—and thus, whether that information constitutes personal data—the GDPR requires controllers to take into account "all the means reasonably likely to be used… either by the controller or another person to identify the natural person directly or indirectly."[104] This includes consideration of factors "such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."[105]

Courts have indicated that effective anonymization is possible, even as they have articulated high and nebulous standards of anonymization. In *Breyer v. Bundesrepublik Deutschland*, for example, the CJEU held that a dynamic IP address in the possession of a website operator might not constitute personal data, even if the same data could be personal data in the possession of an internet service provider, if it would be "practically impossible" to identify the data subject because it would require a "disproportionate effort in terms of… cost, and man-power."[106] The court's analysis focused on whether "legal channels exist" to obtain additional information that would allow an individual to be identified.[107] Not only does this set a high bar for what would qualify as anonymous data, but the resulting standard is highly uncertain in the research context. It may not be possible to know what other information could be available to data recipients to be able to assess the risks of reidentification of a publicly released dataset.[108] Computer science research has demonstrated that the volume of information available online makes this high standard very difficult to achieve.[109]

This highly contextual and fact-specific test involves a case-by-case analysis of the kind that is particularly difficult from the vantage point of a data originator.

---

[104] GDPR, Recital 26.

[105] *Id.*

[106] Case C-582/14, Breyer v. Bundesrepublik Deutschland, *ECLI:EU:C:2016:779 (Oct. 19, 2016)* ¶ 46.

[107] *Id.* at ¶ 47.

[108] In that case, although local law prevented internet service providers from disclosing the identity of an IP address absent a court order, the court found that an IP address could nonetheless be considered personal data since a legal channel existed by which the identity could be obtained (i.e. by obtaining a court order). *Id.*

[109] Luc Rocher et al.,, *Estimating the success of re-identifications in incomplete datasets using generative models*, Nature Commc'ns (2019), *I*.

Guidance from regulators has added further confusion to the standards to apply to the anonymization of personal data. For example, in its guidance from 2014

on anonymization techniques, the Article 29 Working Party stated, on the one hand, that personal data could only be considered effectively anonymized if the process of anonymization was "irreversible," while at the same time stating that an anonymization process "is sufficiently robust" if "identification has become 'reasonably' impossible."[110] These two somewhat contradictory concepts are difficult to apply at the same time and do not create a coherent rule. Moreover, the Working Party's insistence that an anonymization technique is only effective where it is "engineered appropriately," without further detail, only serves to add confusion, since the same technique may succeed in one instance and fail in another, depending on the precise manner in which it is engineered.[111]

These challenges are compounded in the data sharing context because data originators often retain the underlying data (*e.g.* to continue to provide a service to data subjects), while releasing only de-identified data sets. Although the released data sets may not be reasonably identifiable from the perspective of researchers, inconsistent, and sometimes contradictory, guidance at the national level has deepened the confusion. For example, guidance from the Irish Data Protection Commissioner suggests that effective anonymization may not be possible where a data originator "retains the raw data, or any key or other information which can be used to reverse the 'anonymization' process and to identify a data subject," potentially even if researchers could not access this information in the ordinary course.[112] By contrast, joint guidance from the EDPS and the Spanish data protection authority, Agencia Espanola Proteccion Datos ("**AEDP**"), suggests that it may be appropriate to assess whether data is anonymous based on the "likelihood" of re-identification from a party's point of view.[113] In light of these strict and inconsistent standards, data originators have reason to approach anonymization with caution, as even where the possibility of re-identification appears remote, data could nonetheless

---

[110] Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymisation Techniques*, at 8 (Apr. 10, 2014).

[111] *Id.* at 23.

[112] Irish Data Protection Commission, Guidance on Anonymisation and Pseudonymisation, at 5, 7 (June 2019).

[113] Agencia Espanola Proteccion Datos & European Data Protection Supervisor, Introduction to the Hash Function as a Personal Data Pseudonymization Technique (October 2019).

be subject to the GDPR.

b. *Data originators cannot effectively assess the risks of identification where data will be made available to independent researchers.*

Regardless of the appropriate standard to apply, the analysis of whether data is anonymized requires careful consideration of the potential risks of re-identification, taking into account any other information that may be available. The more widely that data will be shared, the more difficult it is for data originators to know and assess re-identification risks. This encourages data originators to aggregate and anonymize data to very high standards so as to reduce these risks. The resulting data often has reduced utility for research. For example, with respect to data concerning social media usage, to limit risks of re-identification, user activities must be aggregated to a sufficient degree that no one individual could be singled out. This may not impair studies of high level trends, but any analysis of more fringe activities—such as highly inflammatory links shared by only a small number of individuals—would be more difficult to assess without compromising anonymity. Researchers criticized the data Facebook shared as part of Social Science One, for instance, in part because Facebook had used differential privacy—an emerging best practice in privacy circles—to help prevent users from being identified.[114] To put a finer point on this difficulty, concerns about anonymity may be valid even where the relevant social media accounts are fake, as a fake social media account could still be linkable to a "natural person."

In the absence of actionable guidance for effectively anonymizing data in the research context, data originators are left to set their own standards. Although this is not inherently problematic from a data protection point of view, leaving key anonymization decisions to data originators could have important implications for the broader research community. The methods and standards of anonymization, even when selected by data originators with the utmost care, could dictate what forms of research are possible. In other words, while data itself is non-rivalrous, the production of privacy-protected datasets can be rivalrous, as the resulting form of the data may enable some studies and preclude others. The development of a code of conduct as proposed by EDMO, which includes broad consultation with relevant stakeholders

[114] Jeffrey Mervis, *Researchers finally get access to data on Facebook's role in political discourse*, Science Magazine (Feb. 13, 2020).

from the research community, could foster anonymization standards that reflect the variety and range of the community's research interests. Developing independent standards of anonymization would also facilitate wider sharing of data, as data

originators could be satisfied that the wider disclosure of such data would not result in unforeseen legal risk. Moreover, if standards of anonymization are clearly articulated, it may not be necessary to impose contractual and other data protection limitations that may compromise research independence.

IV. HOW EDMO'S PROPOSED CODE OF CONDUCT COULD SUPPORT DATA SHARING FOR INDEPENDENT RESEARCH

The GDPR's risk-based approach to research highlights the inherent tension between the public benefits of research and the private rights of individuals. These conflicting imperatives can be heard in statements by EU regulators highlighting "corporate secrecy which characterises the biggest technology companies [as] a barrier to scrutiny by [independent] researchers,"[115] while at the same time commenting that the "close working relationship" between technology companies and researchers can lead "academic studies and the commercial enterprises set up by academics [to] become inextricably entangled."[116] When it comes to big data research, there may be tradeoffs between the independence of researchers, the quality of data they can be permitted to access, and the protection of the rights and interests of data subjects.

Although the GDPR's regime governing research provides an opportunity to carefully balance these interests to optimize research outcomes without unduly compromising privacy and data protection, in practice, because of challenges and inconsistencies in the interpretation of key provisions, the effect of the GDPR's research framework is to encourage stricter protections for personal data, sometimes at the expense of research aims. Indeed, because data originators face significant legal risks if they unduly prioritize research, in the absence of clear standards governing the sharing of personal data for research purposes, data originators are likely to mitigate legal exposure by limiting data sharing and imposing conditions on researchers permitted to access personal data. EDMO's proposal to create a code of conduct for research offers an opportunity to address several of the challenges described above.

1. Potential benefits of developing a code of conduct.

The GDPR encourages the development of codes of conduct that specify the application of GDPR requirements to particular processing activities.[117] Although

---

[115] European Data Protection Supervisor, *A Preliminary Opinion on data protection and scientific research*, at 9 (Jan. 6, 2020).

[116] *Id.* at 7.

[117] GDPR, Art. 40(2).

compliance with a code of conduct does not guarantee compliance with the GDPR, regulators are required to consider a controller's adherence to a code of conduct as *evidence* of compliance with the GDPR and as a factor for reducing the scope of any fines in the event of a GDPR violation.[118] Codes of conduct may also be used to facilitate the cross-border transfer of personal data, which has important implications for the international research community.[119] Codes of conduct, therefore, offer several benefits to data originators and researchers in this context. In addition to the potential for reducing non-compliance risks, codes of conduct also offer the opportunity to articulate clear data protection standards, with particular application to a data provider's industry, which in turn serves to better protect privacy.

Furthermore, relevant stakeholders would have the opportunity to participate in the development of the code's standards, thereby helping to ensure that the resulting standards adequately consider the relevant context. Critically in the research context, codes of conduct require independent bodies to administer the code and enforce compliance. This shifts responsibility for compliance and oversight to the independent code administering entity(ies), which would ensure that the interests of researchers and data originators are taken into account. This shift may help to resolve tensions between the independence of researchers and the accountability of a data sharing regime by positioning an independent third-party to exercise effective oversight.

EU regulators have also expressed enthusiasm for the use of codes of conduct in the research context. In its Guidelines on Codes of Conduct, for example, the EDPB specifically offered examples from the research context as instances where a code of conduct could prove particularly valuable.[120] Codes of conduct also appear in the EDPS Opinion on data protection and scientific research, which emphasized their value in "improv[ing] convergence of practices and increas[ing] confidence in compliance" as well as in "achiev[ing] sufficient levels of harmonisation" across

[118] GDPR, Art. 83(2)(j).

[119] GDPR, Art. 46(1)(e).

[120] EDPB, *Guidelines 1/2019 on Codes of Conduct and Monitoring Bodies under Regulation 2016/679* (June 4, 2019).

EU Member States.[121] Specifically, the EDPS considered that "[s]pecialised codes might be particularly relevant for fields such as biobanking, genomic research or social networks research."[122]

2. Proposed issues to be addressed and relevant stakeholders.

In light of the analysis above, a code of conduct for research should take into account the following aspects:

- **Mission, scope and process:** The code of conduct should have a clearly defined mission and scope. This should include developing standards for the types of researchers and forms of research that would be eligible to receive data pursuant to the code.

- **Data access, anonymization and technical safeguards:** The code of conduct should describe the technical standards to which data should be anonymized or de-identified for varying research purposes and the technical safeguards that would apply to a researcher's access to such data. This may include processes for independent review of research proposals by the code administering entity before access to different tiers of data is granted.

- **Organizational and research safeguards:** In addition to safeguards relating to the data provided by data originators, the code of conduct should include organizational measures that protect privacy, particularly by limiting the potential for misuse after data is shared by data originators. This may include limitations on permitted uses, clear security standards, data minimization and retention limits, risk assessment requirements, and rules governing the further disclosure, international transfers, and publication of information received pursuant to the code. To effectively address the challenges described in this paper, these safeguards should arbitrate the extent to which data originators should exercise oversight and control over research activities

[121] European Data Protection Supervisor, *A Preliminary Opinion on data protection and scientific research*, at 25 (Jan. 6, 2020).

[122] *Id.*

while preserving the independence of researchers. These safeguards should, moreover, be tailored and calibrated to the level of data access provided to researchers and the standards of anonymization or de-identification applied to a given dataset. EDMO and the relevant stakeholders could also consider processes that involve oversight by trusted third-parties, which can be empowered to regulate data access and appropriate safeguards so as to preserve the independence of research activities while satisfying data originators, data subjects, regulators and other interested parties of the compliance of such sharing with privacy laws.

- **Transparency and individual rights:** The code of conduct should also clarify the standards for informing data subjects of the use of data for research purposes, the choices that data subjects will have, and how requests to exercise rights under the GDPR will be addressed.

- **Governance, enforcement and oversight:** Finally, the code of conduct will need to account for the structure of oversight and enforcement bodies as well as the standards for monitoring compliance, receiving complaints, and enforcing compliance.

The development of these standards and processes will require balancing research objectives against the privacy rights of data subjects. To ensure that all relevant interests are adequately protected, the process for developing the code of conduct should include representation from the groups that would be affected by the code, including researchers and research institutions, data originators, and consumer groups. Public authorities with responsibility for data protection and/or the promotion of research may also wish to participate.

V. CONCLUSION

Facebook welcomes EDMO's proposal to develop a code of conduct for social scientific research. EDMO's proposal offers significant opportunities to resolve issues that have led to a situation where we, as a data originator, are disincentivized from sharing personal data with experts and academic researchers in order to ensure that individual data protection rights under the GDPR are protected. EDMO's proposed process, particularly if a code can be developed in consultation with all relevant stakeholders, offers opportunities to:

- Clarify and clearly define the interpretation of GDPR concepts in the area of data sharing for academic research and develop standards that could apply across EU Member States;
- Provide a clear framework that simultaneously ensures that individual privacy interests under the GDPR are protected and makes clear how data originators can share data useful to academic research;
- Ensure the independence of social scientific research through a framework and institutions that would protect privacy without creating incentives for data originators to have direct oversight and control of independent researcher's work; and
- Facilitate an open conversation between relevant stakeholders in the area of social scientific research to help achieve a shared goal of enabling data-driven, privacy-safe, independent academic research that helps us all to better understand the impact of technological developments on society.

We would welcome the opportunity to further engage with EDMO on this worthy initiative.